评分卡模型的前沿研究
—组合评分卡模型

目录

组合模型概述

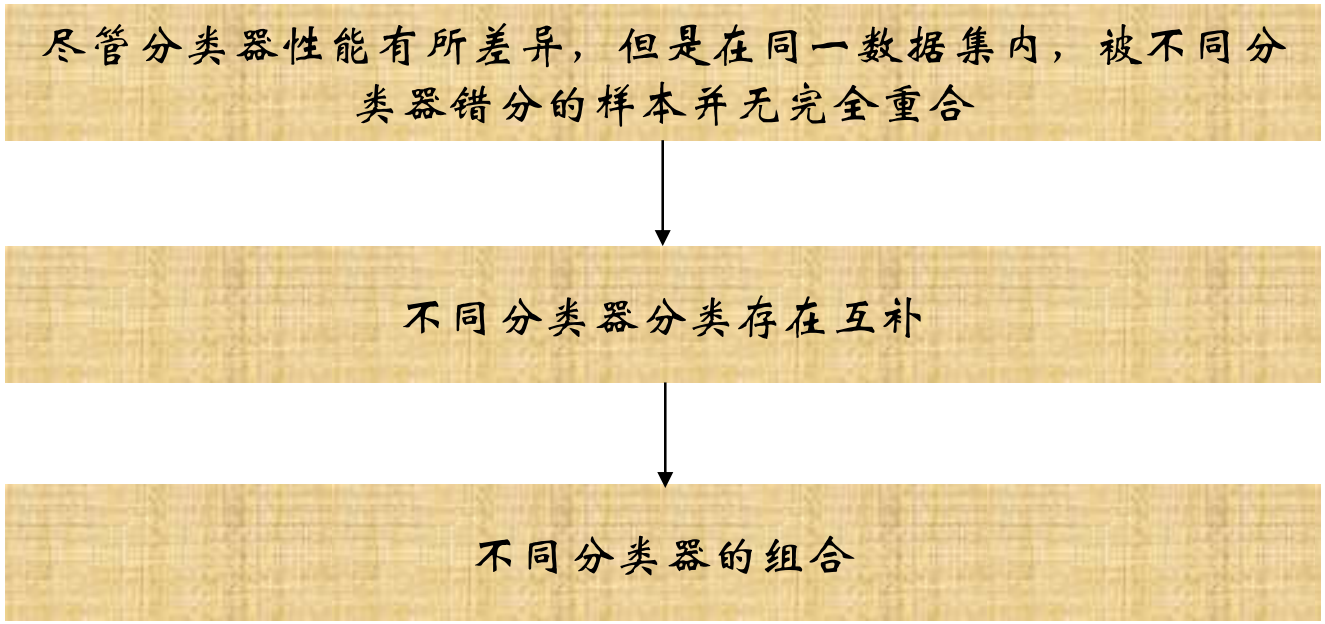
- 什么是组合模型

“把多种单一模型组合起来共同解决一个问题”

- 组合模型的必要性
  - ✓ 能够为评分模型提供更为广阔的发展空间
  - ✓ 能够为信用风险评估的准确性、稳健性最优选择问题给出了答案
  - ✓ 能够提高评分模型的效率

组合模型概述

组合模型的原理



目录

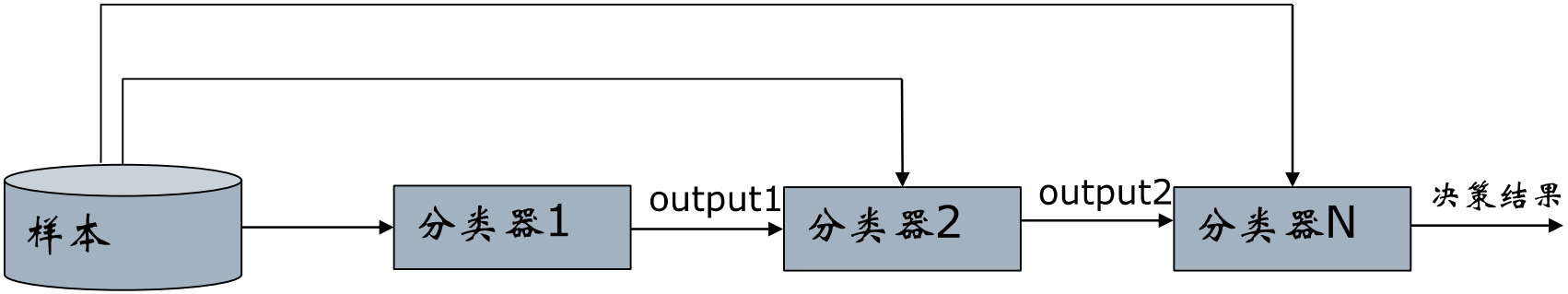
两类结构的评分组合模型

- 串行结构的评分组合模型

- 是指将某单一模型的输出作为另一种单一模型的输入，也成为叠加法
- 在这种类型的组合模型中，即使单一模型的精度不是很高，随着模型个数的增加，分类精度会逐渐提高
- 串行方式的组合模型的性能取决于单一模型的分类精度及组合方式。一般分类精度高的模型排在前面

两类结构的评分组合模型

- 串行结构的评分组合模型(续)



- 串行组合模型的缺点：

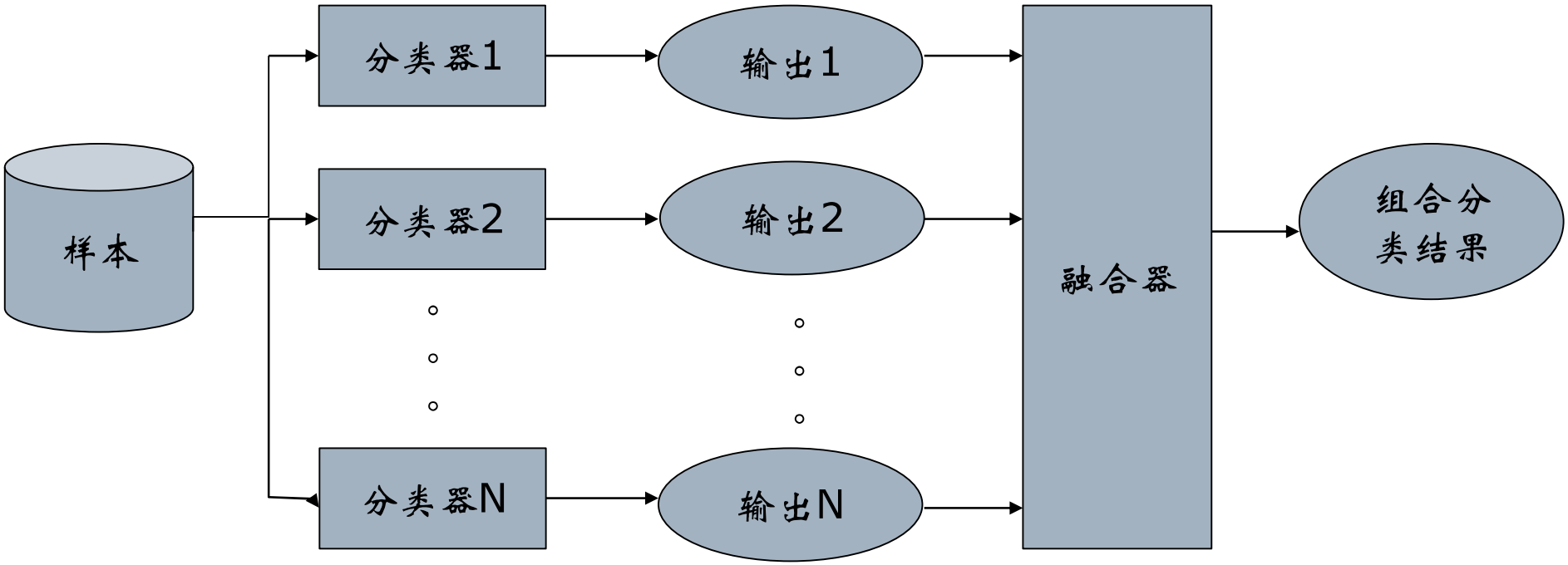
可靠性差。如果有一个分类器出现了错误，错误会传播到下一个分类器，并且可能会放大。

两类结构的评分组合模型

- 并行结构的评分组合模型
- 将多个单一模型(基分类器)的输出通过某种方式进行组合
- 单一模型在构造生成过程中并不依赖其它单一模型，相互间的构造过程完全独立
- 过程可以并行完成

两类结构的评分组合模型

- 并行结构的评分组合模型(续)



与串型结构相比，并行结构具有将强的稳健型。单一分类器的错误不会影响整体的决策

两类结构的评分组合模型

- 混合结构
  - 同时存在串行和并行结构
  - 过于复杂，增加了模型的构建难度
  - 降低了模型的可解释性
  - 不太使用于评分卡模型

两类结构的评分组合模型

- 单一模型的选择

单一模型之间的种类关系可以把组合模型划分为异态组合与同态组合两种。

异态组合

使用不同的分类算法建立单一模型并进行组合

同态组合

使用同一分类算法(参数不同或者建立在不同的训练集上)建立单一模型并进行组合

两类结构的评分组合模型

单一模型的选择(续)

单一模型需要满足以下基本要求

- 单一模型之间的数据或者假设要求要基本相同
- 单一模型的分类错误率要低于0.5
- 单一模型之间要保证相互独立
- 单一模型的复杂度和资料搜集的难易程度也要适度
- 单一模型的数量并非越多越好

目录

组合模型概述

两类结构的评分组合模型

串型组合模型的原理和数值实验

同态并行结构的评分组合模型

串型组合模型的原理和数值实验

串型组合模型的原理

在串型结构的组合模型中，各个基分类器的学习是顺序进行的。后一个基分类器的学习要利用前一个基分类器的学习结果。没有任何一种模型占据绝对优势。

逻辑回归、贝叶斯网络等

- 稳健性好、可解释性高
- 分类精度低于人工智能模型

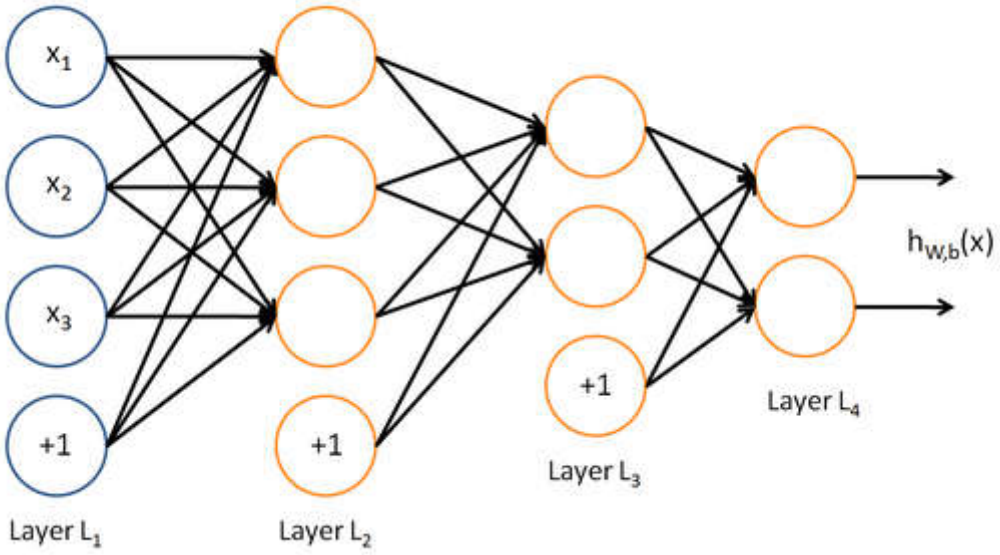
神经网络、支持向量机等

- 分类精度高
- 缺乏稳健性和可解释性

串型组合模型的原理和数值实验

串型组合模型的数值实验

在原始变量完成分箱和WOE编码后，我们随机选取了20个变量构建了神经网络模型。该模型含有2层隐藏层，第一层含有5个节点，第二层含有2个节点。该模型没有实施最佳调优。



串型组合模型的原理和数值实验

串型组合模型的数值实验(续)

得到神经网络模型的结果后，将预测的概率作为逻辑回归的一个输入特征，和其它WOE编码后的特征一起构建逻辑回归的模型，要求依然是回归模型的系数为负(除了神经网络输出的概率外)且显著。

该模型和单一的逻辑回归模型相比，在训练集上的表现有一定的提升

	KS	AR
单一逻辑回归	36.0%	73.7%
神经网络＋逻辑回归	36.2%	73.8%

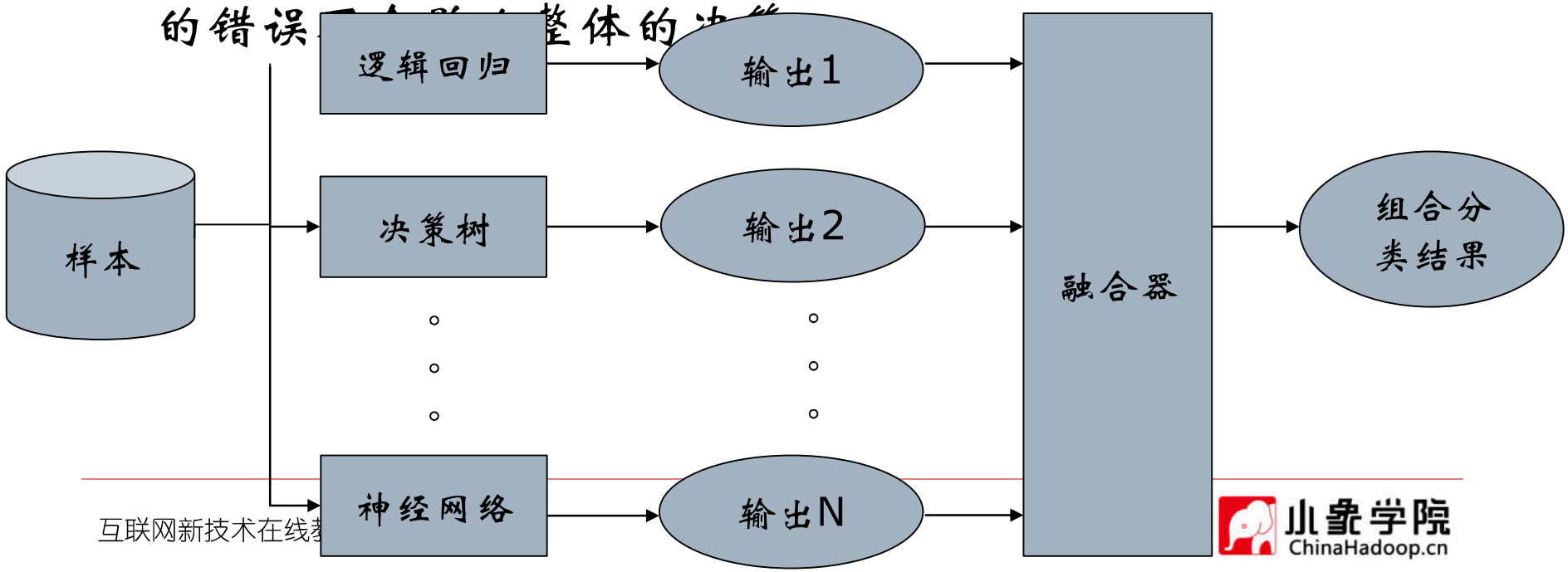
目录

并行组合模型的原理和数值实验

- 异态并行结构的信用评分组合模型的原理

由于不同分类算法在原理上存在很大的差异，因此通过选择不同的算法来构建不同的单一模型并组合是很自然的思路。

与串型结构相比，并行结构具有将强的稳健型。单一分类器的错误[illegible]整体的[illegible]



并行组合模型的原理和数值实验

- 异态并行结构的信用评分组合模型的原理(续)

在并行模型的融合阶段，需要按照某种策略对基分类器的结果进行合成以获得最佳分类效果。

- 投票法(输出结果是0/1)
  - 简单投票
  - 加权投票
- 代数合成法(输出结果是连续数值)
  - 平均值法
  - 加权求和法

并行组合模型的原理和数值实验

- 异态并行结构的信用评分组合模型的数值实验

在串型组合模型的实验里面，我们随机挑选了20个变量构建了未调优的神经网络模型，得出了预测概率。将预测概率转换成log odds，再和逻辑回归评分卡模型给出的log odds求加权平均，得出的组合性的log odds的表现如下：

	KS	AR
单一逻辑回归	*36.0%*	*73.7%*
神经网络与逻辑回归并行	*36.2%*	*74.0%*

并行组合模型的原理和数值实验

- 同态并行结构的信用评分组合模型的原理

- 异构并行的组合模型虽然精度高，但是由于采用不同算法的模型，易增大模型构建的复杂性
- 可以采用相同的分类算法，在不同的样本上训练基分类器产生差异性，再进行组合

两种思路

- 采用同样的分类算法在不同的样本上训练分类器，如Bagging，Boosting
- 采用同样的分类算法在不同的特征空间上训练分类器，如RSM

并行组合模型的原理和数值实验

- Bagging算法

基本思想

对训练集进行有放回的抽样，获得多个不同的训练子集，再采用某种学习算法在不同的训练集获得多个具有较大差异的基分类器，再组合起来

注：

这种基分类器必须是不稳定的，常用的算法有决策树、神经网络、支持向量机等

并行组合模型的原理和数值实验

Boosting算法

基本思想

把若干个弱分类器(即预测精度略高于随机预测的分类器)组合起来提升为强分类器。

特点

- Boosting算法中的基分类器是串型生成的，每个基分类器在训练集上的误分情况被用于调整下一个基分类器的训练集
- 最后加权投票把每个基分类器的输出结果进行组合。

并行组合模型的原理和数值实验

- RSM(Random Subspace Method)算法

基本思想

通过随机选取特征集的办法来产生不同的训练集，因此能够产生具有足够差异性的基分类器。

特点

- Boosting算法中的基分类器是串型生成的，每个基分类器在训练集上的误分情况被用于调整下一个基分类器的训练集
- 最后加权投票把每个基分类器的输出结果进行组合。

并行组合模型的原理和数值实验

- Bagging算法的数值实验

第一步，我们从训练集里有放回地抽取60%的样本和5个特征构成一个训练集，构建神经网络模型，再在测试集上预测概率和log odds

第二步，重复第一步9次，共计得到10个预测的log odds，求出平均值

结果显示，这种方法得出的模型的表现，好于单一神经网络在测试集上的表现

	KS	AUC
单一神经网络	*34.6%*	*73.4%*
Bagging	*35.6%*	*73.70%*

疑问

- 小象问答官网
  - http://wenda.chinahadoop.cn

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象学院
- 新浪微博：小象AI学院